Supervised projection pursuit - A dimensionality reduction technique optimized for probabilistic classification

Barcaru, Andrei

Published in:
Chemometrics and Intelligent Laboratory Systems

DOI:
10.1016/j.chemolab.2019.103867

Chemometrics and Intelligent Laboratory Systems 194 (2019) 103867

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemometrics

Supervised projection pursuit – A dimensionality reduction technique optimized for probabilistic classification

Andrei Barcaru

Department Laboratory Medicine, University of Groningen, University Medical Center Groningen, 9700 RB, Groningen, the Netherlands

A B S T R A C T

An important step in multivariate analysis is the dimensionality reduction, which allows for a better classification and easier visualization of the class structures in the data. Techniques like PCA, PLS-DA and LDA are most often used to explore the patterns in the data and to reduce the dimensions. Yet the data does not always reveal properly the structures wen these techniques are applied. To this end, a supervised projection pursuit (SuPP) is proposed in this article, based on Jensen-Shannon divergence. The combination of this metric with powerful Monte Carlo based optimization algorithm, yielded a versatile dimensionality reduction technique capable of working with highly dimensional data and missing observations. Combined with Naïve Bayes (NB) classifier, SuPP proved to be a powerful preprocessing tool for classification. Namely, on the Iris data set, the prediction accuracy of SuPP-NB is significantly higher than the prediction accuracy of PCA-NB, (p-value $\leq$ 4.02E-05 in a 2D latent space, p-value $\leq$ 3.00E-03 in a 3D latent space) and significantly higher than the prediction accuracy of PLS-DA (p-value $\leq$ 1.17E-05 in a 2D latent space and p-value $\leq$ 3.08E-03 in a 3D latent space). The significantly higher accuracy for this particular data set is a strong evidence of a better class separation in the latent spaces obtained with SuPP.

1. Introduction

Dimensionality reduction (DR) techniques, also referred to as projection methods, are perhaps the most used exploratory tools for applications in various fields, from image analysis and information retrieval to bioinformatics and chemometrics. The main reason for such an extensive use of these techniques is the set of benefits that these are bringing for data analysis: (i) possibility to plot and visualize data and potential structures in the data in lower dimensions, (ii) possibility to apply stochastic models, (iii) capacity to solve the "curse of dimensionality" and (iv) facilitation of prediction and classification of the new data sets (i.e. query data sets with unknown class labels). The projection techniques can be classified into three major groups according to the way the latent components are obtained: supervised (i.e. considers class labels for the deduction of the latent components and for further classification), semi-supervised (i.e. uses both labeled and unlabeled samples to infer class structures in the latent space) and unsupervised (i.e. class labels are not available and are yet to be found from the structural patterns of the projected data or the class labels are simply not used). Each of these three major types of DR methods can be further divided into "linear" and "non-linear" methods. A myriad of methods emerged in the past two decades, and listing all of them would be a task for a detailed review. We ought to mention a few however from each category that are more recent or more used across different domains.

From the unsupervised category and the subcategory of linear methods, the most commonly used are principal component analysis (PCA) [1,2], projection pursuit (PP) [3–5] and independent component analysis (ICA) [6]. In the non-linear subcategory, it is worth mentioning the Kernel PCA (KPCA), local linear embedding (LLE) and isometric mapping (Isomap) [7]. The semi-supervised methods are a relatively new form of DR. Hou et al. described a multiple view semi-supervised DR (MVSSDR) [8] and more recently, Mikalsen et al. introduced a noisy multi-label semi-supervised DR (NMLSDR), which solves the problems of missing labels and noisy labels [9].

From the supervised methods, the most used are partial least squares (PLS) [10,11], orthogonal projection to latent structures (OPLS), including the combination with discriminant analysis of both (i.e. PLS-DA and OPLS-DA), Fisher's discriminant analysis or linear discriminant analysis (LDA) [12], heteroscedastic discriminant analysis (HDA) [13], regularized discriminant analysis (RDA) [14] and the regularized coplanar discriminant analysis (RCDA) introduced by Huang et al. [15]. In the non-linear subcategory of the supervised DR, Örnek and Vural [16] recently introduced the smooth regular embedding (SRE) and Raducanu and Dornika proposed a supervised version of the linear embedding DR (SLE) [17]. Pires et al. [18,19] and Lee et al. [20] described a combination between PP and LDA thus including the PP technique also in the group of the supervised projection techniques.

There is a considerable amount of research directed on comparison between non-linear DR methods and the linear ones [21–24]. Most of them outlined that the non-linear DR techniques are much slower than

E-mail address: a.barcaru@umcg.nl.

https://doi.org/10.1016/j.chemolab.2019.103867
Received 15 January 2019; Received in revised form 2 October 2019; Accepted 6 October 2019
Available online 9 October 2019

the linear ones and that the non-linear methods are having a better performance than the linear ones on synthetic data sets and poorer performance on real data sets. It must be added here that one of the major concerns with the non-linear methods is the interpretability of the latent space (i.e. the latent manifold). In the linear DR methods, a component can express the percentage of variability for example between various species of flowers and/or the variance within one specie (e.g. as the principal components in PCA and OPLS). Sometimes the latent component can express the direction of the maximal effect size, as in the case of LDA. It is the interpretability that gives us the possibility to extract certain important variables by the degree to which these are contributing to the interpreted property. Such variable selection procedures are the magnitude of the loadings in the PCA and so-called variable importance on projection (VIP) in the case of PLS (and OPLS). On the other hand, the eigenvalues of a manifold or the components obtained with a non-linear DR are much harder to interpret for a real world example. However, the linear methods have their own shortcomings. Namely, in order for the LDA to be properly applied the projected data have to comply with homoscedasticity and normality. This may be a non-realistic requirement especially for the biological data where outliers and biological variability are often present. Several options emerged to solve these limitations of LDA among which are the HDA and Multimodal Oriented Discriminant Analysis (MODA). The latter is based on the maximization of the pairwise, symmetric version of Kullback-Leibler (KL) divergence between the distributions of the classes. Abou-Mustafa pointed to the limitation of the total pairwise KL divergence maximization approach and further extended MODA to Pareto discriminant analysis (PARDA) [25]. PARDA is based on Pareto multi-objective optimization considering all pairwise KL divergences at the same time. Although an elegant solution, the Pareto optimization strategy is not necessary when a different divergence is employed, namely Jensen-Shannon divergence, which is the core of the method proposed in this article. Another limitation of LDA is the dependency between the number of classes and the number of latent components, i.e. the latter has to be strictly smaller than the number of the classes, unlike PCA or PLS, where maximum number of components is equal to the number of attributes. Recently, Gromski et al. published a review pointing to the weaknesses of PLS-DA, suggesting the use of alternative methods such as Principal Component-Discriminative Feature Analysis (PC-DFA) [26]. Several serious disadvantages of PLS-DA were also previously reported in Ref. [27]. For example, when the number of sample points is considerably smaller than the number of attributes (i.e. variables), PLS-DA can give a good classification just by chance (i.e. overfitting).

In this article is presented an adaptation of projection pursuit to a supervised way of projection analysis by means of entropic divergence measurement, namely Jensen-Shannon divergence. The technique will be referred to as SuPP (i.e. the acronym for Supervised Projection Pursuit) for convenience. The benefits of the SuPP are: (i) maximization of the distances between all the classes simultaneously on each projection, (ii) determination of latent components using a non-parametric approach, (iii) minimization of the Bayes classification error and (iv) capability to handle missing data.

The performance of the proposed strategy is assessed using real data sets at different training-to-test set ratios. Comparative assessment is made using the most applied dimensionality reduction techniques from the same category (i.e. linear DR): PCA, PLS-DA and LDA. The assessment also includes the capability of working with missing data.

2. Supervised projection pursuit

2.1. Entropic projection index

In 1973, Friedman and Tukey [3] explored the idea of Kruskal [28] and outlined a strategy for PP. The core of their method is the search of the projections based on a so-called projection index (PI) that would quantify the "usefulness" or "interestingness" of the projections. Later, Huber [29] generalized the PI and was the first to propose standardized negative Shannon entropy as a PI. Huber's definition of "interestingness" of a projection is linked to the concept of normality of the projection. Huber stated that the more convoluted are the projected clusters, the more normal is the projection and thus less interesting. However, Huber also mentioned that the entropy is by far not the only possibility for a PI and listed a number of possible PIs including Fisher's information and kurtosis. Jones and Sibson [30] proposed to use kernel density estimation of the projected data and, using Rényi generalization of the α-order entropy [31], further shaped the choice of entropic index into an even more rigorous mathematical form. To be more specific, let's consider the projection of the vector $X = [X_1, X_2, \ldots, X_M]$ onto a unit vector $\widehat{k} = [k_1, k_2, \ldots, k_M]$:

$$\chi = X \cdot \widehat{k} \tag{1}$$

Here χ is the set of projected points onto direction vector $\widehat{k}$. Huber's entropic index, according to Jones and Sibson, takes the following form:

$$g(\chi) = \int \widehat{f}(\chi) \ln \widehat{f}(\chi) d\chi \tag{2}$$

where $\widehat{f}(\chi)$ is the kernel density estimation of the projected data χ. The PI from eq. (2) can take positive and negative values and as such cannot be considered an information metric.

Shannon entropy quantifies the information of a system, and is calculated using discrete probability distribution:

$$H(\chi) = -\sum_i p_i \log_2 p_i \tag{3}$$

where $p_i = p_i(\chi)$ is the probability distribution of a discrete random variable χ. The notation $H(\chi)$ can be sometimes written as $H(p)$.

The aim of SuPP is to find the projection that would yield the maximum information from the projected data with the use of group labels and Shannon entropy (SE).

2.2. Supervised adaptation using Jensen-Shannon divergence

In a supervised method, the labels of the classes are available and the "interesting projection" is defined as the projection that yields the best class separation. In other words, the distances between the classes are maximized for a maximal PI. This can be achieved using Jensen-Shannon divergence defined as follows. Let's consider the class labels $Y_C = [y_{11}, y_{21}, \ldots, y_{jc}, \ldots, y_{NC}]$, $c = 1, 2, \ldots, C$ coming from a total of C classes and associated to each data point in the vector X. In this case, eq. (2) takes the form of the SE of the mixture of C class distributions. Subtracting the weighted summation of the entropies for each class distribution from the entropy of the mixture distribution, we obtain the Jensen-Shannon divergence (JSD):

$$D_{JS} = H\left(\sum_c \pi_c p(\chi|c)\right) - \sum_c \pi_c H(p(\chi|c)) \tag{4}$$

Here, $\pi = [\pi_1, \pi_2, \ldots, \pi_C]$ is the vector of weights associated to each class distribution. $p(\chi|c)$ is the discrete probability distribution of the projected data associated to the class c and is calculated using the histograms of the projected data corresponding to a class c. Unlike the differential entropy that can take both positive and negative values, SE can take only positive values and consequently has convenient properties (e.g. is a metric of information, Jensen's inequality can be applied etc). However, as shown in eq. (3), SE is only applicable on discrete variables. In order to accommodate for both continuous and discrete variables, in SuPP a discretized kernel density estimation (KDE) is used. More specifically, a kernel density is obtained using Gaussian KDE and further thinly discretized and normalized to the summation of all the discrete values. Employing the discretized KDE would result in a smoother

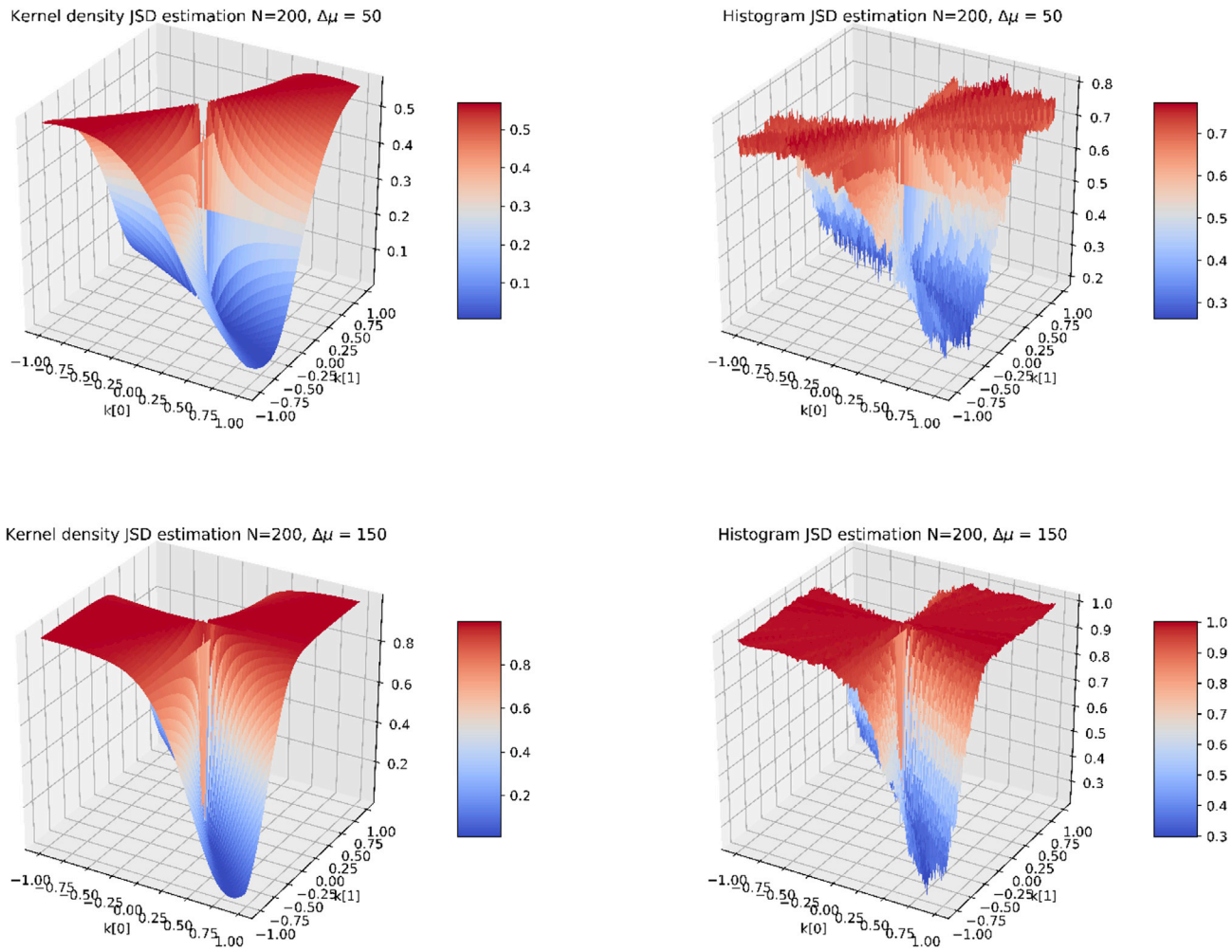

Fig. 1. Two ways of calculating JSD for a data set projected on a rotating vector: using Gaussian KDE (left), using histograms (right). The upper two plots depict the calculations applied for a lower difference between the groups ($\Delta\mu = 50$). Lower two plots are representing the same 2 algorithms applied on a larger difference between the groups ($\Delta\mu = 150$).

objective function (Fig. 1). In the example from Fig. 1, two simulated classes were used to illustrate the effect of the KDE of the probability density function. The centroids of the sample groups were spread apart from one another at a smaller distance (the upper subplots) and at a higher distances (the lower plots). The two subplots on the right represent the rough JSD surface obtained using only a histogram for the estimation of the probability density function. The left subplots show the smoothing effect of the application of the KDE for the calculation of JSD.

JSD is a distance metric that measures the divergence between distributions and therefore is a good choice for a PI [32] when the maximal separation between the projected class distributions is the defining criterion of the latent vector's orientation. This order-1 divergence [33], is particularly interesting when the main aim is to preprocess the data for further probabilistic classification. In fact, Lin showed in Ref. [34] the link between JSD and Bayes error of classification $P(e)$. Lin pointed to the fact that the upper bound of the Bayes error has the following expression:

$$P(e) \leq \frac{1}{2}\left(H(\pi) - D_{JS}\right) \tag{5}$$

Here, $H(\pi)$ is the entropy of the prior probabilities for each class. For equal weights and uniform prior probability assumption $p(c)\tilde{}U(0, C)$, the weights are equal to the prior probabilities. For this work, only the equal weights or equal priors on the classes are used. The goal of SuPP is to find the projection that maximizes the divergence between the

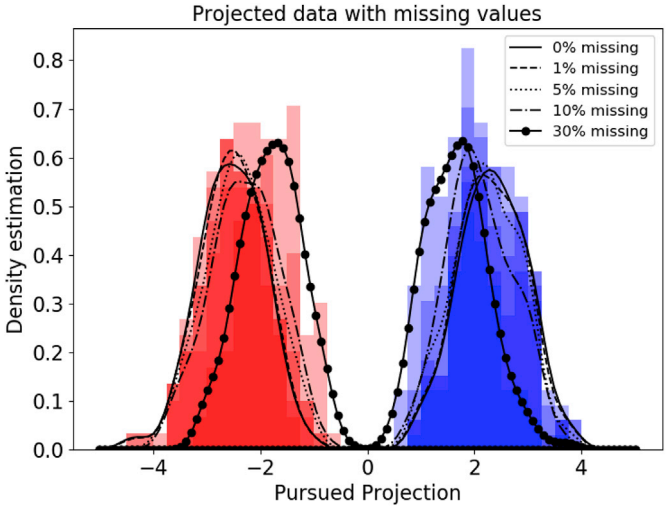

Fig. 2. The histograms and the KDE for each simulated sample group at different levels of missing observations.

probability density functions of the projected data associated to each class (i.e.$p(\chi|c)$) while minimizing the upper bound of the Bayes error. From eq. (5) it fallows that both objectives can be achieved by

```
SuPP(X,Y,D,parameters){
    k ← zeros(rows=D,col=N) //initializing k
    // parameters contain N,D,C,w,π,beta,gamma
    for d in [1:D] {
       start_point ← uniform_random(raw = 1, col = N,limits = [0,1])
       k[d,1:N] ← optimize(ObjectiveD,X,k,parameters,start_point) //Basin-hopping call
       }
    return k
}

ObjectiveD (d, parameters,k, k_proposed){
       π ← parameters.π //group weights e.g. 1/C
       w ← parameters.w //the bandwidth of the KDE, e.g. "scott" or a number
       alpha ← parameters.gamma
       beta,gamma ← parameters.[beta,gamma]
       F ← 0 //initializing mixture distribution
       Hf ← 0 //initializing weighted cumulative entropy
       Orth ← 0
       C ← parameters.C //get the number of classes from the parameters
       for c in [1:C]{ //iterating over groups
              projected ← X[:,class_index == c]·k_proposed
              f ← KernelDensity(projected,w)
              F ← F+π[c]*f
              Hf ← Hf + π[c] * -trapz(f*log2(f))
       }
       HF ← -trapz(F*log2(F)) //calculating the entropy of mixture distribution
       DJS ← HF - Hf
       if d>1{ // calculate the orthogonality term
              O ← 0
              for i in [1:d-1]{
                     for j in [i+1:d]{
                            O ← O + abs(sum(k[i,:]*k[j,:]))
                     }
              }
              Orth ← 2*O / (d(d-1))
       }
        J ← (1-DJS/log2(C))^alpha + gamma*Orth^beta
       return J
}
```

Fig. 3. The SuPP algorithm. The function SuPP makes use of an optimization function "optimize" which aims for an optimal result from ObjectiveD. The SuPP makes use of the training data set X and the class labels Y the number of dimensions of the latent space D the set of "parameters".

maximizing D_{JS}.

The KDE estimation has another beneficial effect on the projected data χ – smoothing of an incomplete histogram. This effect is notable when there are missing values in the data set. Fig. 2 illustrates the histograms and the estimated densities at different rates of missing values (i.e. 0%, 1%, 5%, 10% and 30%). The values were extracted from random locations in the data set and the pursued projection is the one that maximizes the JSD. Note that below 30% the estimated densities do not vary significantly from the 0% missing value rate. At the 30% rate of missing values, the distributions of the two simulated classes appear closer to each other. For a good separation between the classes, this effect plays a minor role on the JSD calculation and the latent component would be close to the one obtained in the rest of the missing value cases.

2.3. The optimization function and the SuPP algorithm

The latent components in SuPP are found sequentially, a strategy suggested also by Friedman and Tukey in the unsupervised PP. By using the PI from eq. (4) in an objective function, one can employ an optimization algorithm to find the latent component that would maximize the distances between the class distributions. Jones and Sibson mentioned in their work that the optimization criteria for the PP are:

(1) Achieving maximum PI (in this case maximizing the divergence D_{JS})

(2) Finding the most orthogonal projection space

Previous publications on PP suggested that it is worth looking for several interesting projections. This aspect of PP is considered, in fact, its strength as opposed to other dimensionality reduction techniques which have only one solution per component. Furthermore, Huber stated in his publication that "orthogonal directions do not suffice, the interesting directions may be oblique to each other" [29]. This suggests, on one hand, that a sequential direction-wise pursuit is preferred to the projection on a multidimensional orthogonal space. On the other hand, the same statement suggests that a penalty coefficient may be used on the second objective listed above, namely on the orthogonality (i.e. to allow the exploration of oblique components). Using these criteria, the optimization function proposed here, is:

$$J(D_{JS}, O) = f\left(D_{JS}^{\alpha}\right) + \gamma f(O^{\beta}) \tag{6}$$

Here, $f(D_{JS})$ represents a function of variable D_{JS} described in eq. (4), the $f(O^{\beta})$ is a function of the average absolute pairwise orthogonality O:

$$O = \frac{2}{D(D-1)} \sum_{d=1}^{D-1} |\hat{k}_d \cdot \hat{k}_z| \tag{7}$$

with $d < z \leq D$, $d \in [1, D-1]$, $z \in [2, D]$ are the indices of the latent vectors corresponding to one dimension and D is the dimension of the

final latent space on which the data is projected. The coefficient γ is meant to weigh the importance of the orthogonality. If this parameter would be $\gamma = 0$ for $D > 1$, the optimization function may return, as the most optimal solution, the same latent component as in the case of $D = 1$. In higher dimensions, γ allows for more oblique components (for lower values of γ) or more orthogonal components (for higher values of γ). Intuitively, there is no need for the orthogonality term when applied in unidimensional space, thus for $D = 1$, $\gamma f(O^{\beta}) = 0$. Note that D_{JS} and O both depend on the unit vector $\widehat{k}$ through the projected vector χ.

The optimal $f(D_{JS}^{\alpha})$ is achieved when D_{JS} approaches its upper limit ε (i.e. achieving maximum divergence between the class distributions). Following Lin's work [34], for equal class weights, $\varepsilon = \log_2 C$. For the case where the entropy terms in eq. (4) are calculated using natural logarithm, the logarithm in the expression of ε changes into a natural logarithm.

Accounting for all the aspects mention above, a good choice for the optimization function is the following:

$$\begin{cases} J(D_{JS}, O) = \left(1 - \dfrac{D_{JS}}{\varepsilon}\right)^2 + \gamma O^2 & , for\ D > 1 \\ J(D_{JS}) = \left(1 - \dfrac{D_{JS}}{\varepsilon}\right)^2 & , for\ D = 1 \end{cases} \quad (8)$$

For an extensive exploration of different local minima, algorithms like genetic algorithm and global search are proposed in Ref. [5] over a simple gradient descent optimization. For the case of large variables-to-samples ratio, Hou and Wentzell suggested in Ref. [35] a regularized objective function for an unsupervised PP regression. For SuPP, a Monte-Carlo based global optimization algorithm developed by Li and Scheraga (also known as "basin-hopping algorithm" or BH) was used [36]. A succinct description of BH algorithm is as follows.

Step 1. Generate a set of points at random in the variable space (Xi)
Step 2. Apply a local search algorithm to generate a local minimum (Yi)
Step 3. Apply a perturbation to Yi to get a new Xi+1 and re-apply the local search on the Xi+1
Step 4. If Objective (Yi+1) < Objective (Yi), retain Yi+1 and increment i

Steps 3 and 4 are applied in a while loop until a stopping criterion is met (e.g. reaching a limited number of iteration).

The local search method or local minimization algorithm used in the steps listed above is based on linear interpolation modelling of the objective function (i.e. COBYLA) developed by Powell in 1994 [37]. The Objective mentioned in the 4th step is the objective function expressed in eq. (8). The reason behind the choice of these two algorithms is, for one part, the extensive (local and global) search capability of the BH algorithm and for the other part, the capacity of COBYLA to work without derivatives. The latter property is required for SuPP due to non-parametric nature of kernel-estimated densities.

For python 3.0 and higher, both algorithms, COBYLA and basin-hopping, are available in the SciPy (https://www.scipy.org/) framework [38]. The pseudo-code of the SuPP algorithm is shown in Fig. 3. Examples of usage and detailed explanations of the SuPP class are provided in section 3 of the supporting information. The script is available at https://github.com/ABarcaru/SuPP under the Apache v2 license.

2.4. Starting point of the optimization algorithm

As many other optimization procedures that have more than one solution, SuPP is, to some extent, sensitive to the starting point. Friedman and Tukey and later Jones and Sibson suggested that using the first principal component can be a good option for a starting vector. However, this may lead to the convergence to the same local minimum which, in case of PP technique, is not always desired as was mentioned previously. Starting with a random vector k increases the chances of exploring new "interesting" projections. Different values of γ can also lead to different and potentially interesting projections in the higher dimensions.

3. Materials and methods of validation

3.1. Objectives

For assessment of the SuPP several objectives must be analyzed: the visualization and exploratory potential, optimization of the parameters used by SuPP and the classifier attached, capacity to handle missing data and the assessment of the quality of separation between the classes (i.e. as pre-processing for classification).

Different data sets are used for the purposes aforementioned.

3.2. Data

A. Synthetic data sampled from normal distribution consisting of two groups "Control" and "Case" having 200 variables. The data was duplicated and a difference in mean values was added to 20 randomly chosen variables to simulate difference between the groups. The structures of the synthetic data sets are listed in Table 1. The data set A.I is only used to generate data A.II and A.III.
B. "Iris data set" or "Fisher's Iris data set" collected and first published in 1935 by Edgar Anderson [39] and later used by Fisher in application of linear discriminant analysis and published in Annals of Eugenics [12]. This data set is made available in *scikit-learn* library [40]. Iris data set has 150 sample points, consisting of 4 attributes (i.e. sepal length, sepal width, petal length and petal width) and 3 cluster labels (i.e. *Iris virginica*, *Iris setosa* and *Iris versicolor*). This data set is typically used by the machine learning community in testing classification, variable selection, projection and dimensionality reduction algorithms.
C. Wine data set, also available in the *scikit-learn* library. This data set has 178 sample points and 13 attributes and 3 class labels [41].
D. Adenocarcinoma data published and made available by Notterman et al., in 2001 [42], which consists of gene expression values obtained with mRNA assay. This data set contains 7500 variables (i.e. genes) and 36 sample points split into two classes: Control with 18 sample points and Tumor with 18 sample points respectively. The adenocarcinoma data set is available in its transformed form as described in the original publication of Notterman et al. (i.e. housekeeping gene normalization, replacement of ND with 0).

3.3. Methods of validation

3.3.1. Data exploration capacity

The exploratory data analysis is based on the visual inspection of the projected data in lower dimensions. The potential of a projection method as an exploratory tool lies in the quality of separation of the classes on the

Table 1
Three synthetic data sets.

	$\overline{\Delta\mu}$	σ_w	σ_b	N_T	N_V	M
A.I	0	10,10	0	250	250	100
A.II	30	10,15	15	250	250	100
A.III	30	10,15	15	25	250	100

A.I – no difference between the mean of the groups ($\Delta\mu$), with the same within-group standard deviation (σ_w) and with the same number of training and validation points (i.e. N_T and N_V respectively).
A.II - added difference between the groups, increased the within-group standard deviation of the group "case".
A.III – same as A.II with the exception of the number of training points (i.e. reduced to 25).

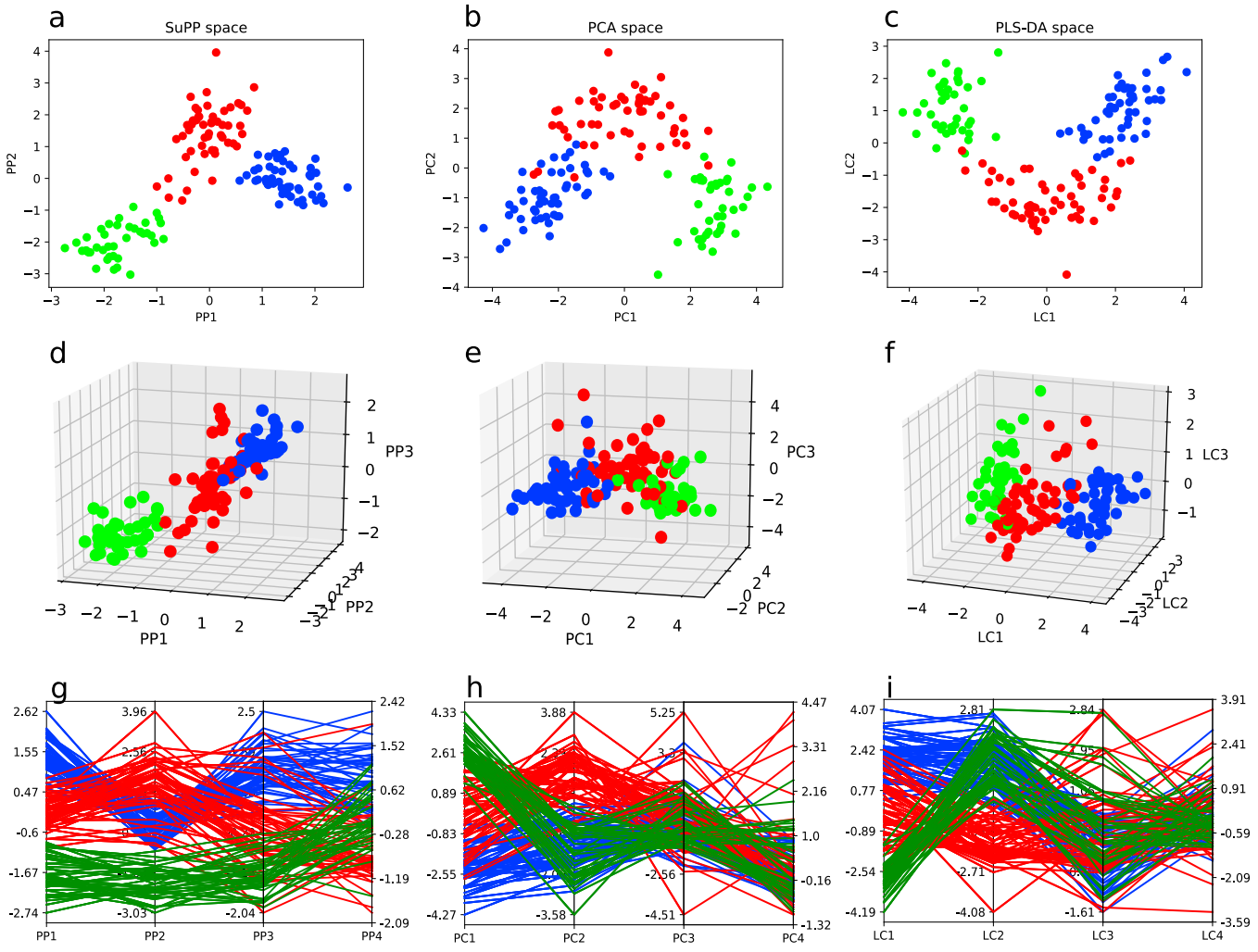

Fig. 4. The visualization of the Wine data set. SuPP (a, d, g), PCA (b, e, h) and PLS-DA (c, f, j). The lower subplots (g, h, i) are representing the data in parallel coordinates.

latent components. Visual estimation of the separation of the classes in the projection space is done typically by plotting the projected data onto a 2D and 3D spaces. However, a good visual estimation of the separation of the classes can be achieved through the plotting using "parallel coordinates". In case of good separation, the classes are well stratified at least on one axis from the parallel coordinates plot. For a linear relationship between the classes and the latent components, the stratification must be observed on all the axes. For the visualization and exploratory potential evaluation of SuPP, data set C is chosen. To quantify the quality of separation of classes, the following metrics were employed: Mahalanobis distance, Bhattacharyya distance and Hellinger distance. Detailed description of each metric is available in the SI, section 1. Each distance was calculated pairwise: Group 1 - Group 2 and Group 2 – Group 3. Statistical significance of the pairwise separation between the groups was estimated using Hotelling's T^2 and F-statistics.

3.3.2. Parameter optimization

Prior to the comparison of SuPP-SVM and SuPP-NB with other techniques, a few parameters must be optimized. More specifically, for SuPP, the penalty on the orthogonality, i.e. γ (eq. (6) and eq. (8)), and for SVM, the influence of a sole training sample (i.e. Γ) and the maximum decision function's margin, i.e. C_{DFM}. The data sets A.II and A.III are used for the optimization of the parameters (i.e. γ, Γ and C_{DFM}). The penalty γ is drawn from the powers of the number of classes (C). The optimality of γ is indicated by the accuracy of the Naïve Bayes classifier at different levels of γ. In order to optimize the number of Monte Carlo iterations (i.e. ***niter***) of the basin-hopping algorithm, this parameter is increased with different levels (i.e. 100, 500, 700 and 1000) and the accuracy is evaluated 5 times at each increment. For the optimization of the SVM classifier, the accuracy of the prediction of this classifier is assessed at $\Gamma = [0.1, 1, 10]$ and for each value of Γ, $C_{DFM} = [0.01, 1, 100]$.

Another parameter could be optimized, namely bandwidth of the KDE. However, for this work, the default method of the kernel density bandwidth calculation is used, i.e. $bw = nef^{-\frac{1}{d+4}}$ with *nef* representing the number of effective points and $d = 1$ the dimension of the projected data.

3.3.3. Evaluation of the capacity to handle missing data

To illustrate the capacity of SuPP to work with missing values, the synthetic data A.II was used. In the training set, a percentage of the total values was replaced with "NaN" (i.e. 1%, 5%, 10% and 30% missing values of the entire training data set) at different random locations, similar to the case indicated in Fig. 2.

3.3.4. Assessment of SuPP as pre-processing for classification

In order to assess the performance of SuPP as a pre-processing technique aimed for further classification, the data is split into two sets: training set and test set. The training set is used to obtain the latent components and classification models. The test set is projected on the latent components obtained with the training set and the classification

model is applied on the projected test set. The accuracy of the classification of the test data set reflects the potential of the projection method. As classifiers, Support Vector Machine (SVM) and Naïve Bayes (NB) are employed due to their superiority over DA or logistic regression classification as mentioned in Ref. [43]. The Radial Basis Function kernel is chosen for the SVM for its robustness to non-linearity in the data.

For this purpose, data sets B, C and D are used. The SuPP-SVM and SuPP-NB are compared with PCA-SVM, PCA-NB, PLS-DA and LDA by comparing the accuracies of the prediction of the test set. To assess how significant is the prediction accuracy, we must first obtain distributions of accuracies. To this end, the method used for testing the quality of separation between the classes is based on cross-validation. More specifically, the algorithm iterates over 3 variables: (i) the number of components (ii) the fraction of data set splitting (i.e. ratio of the test sample size to the training sample size), and (iii) the random state (Fig. S1). For the number of components, maximum of 3 were selected thus exhausting the maximum number of dimensions for visual representation of the complex data. If needed, this number can be further optimized (i.e. increased or decreased) with a double cross validation technique reported in Ref. [44]. The second loop iterates over the ratio between the test set size and the training set size. The algorithm iterates over the range 0.16–0.75 with 5 equidistant values in total. The limits were selected to ensure that the data set containing the smallest number of sample points, in this case adenocarcinoma, has at least 3 points for the test and training respectively. The last loop, having 20 iterations (minimal number recommended for Mann-Whitney *U* test), ensures the random selection of the data points (i.e. different topology of the points in each random sampling) yielding a distribution of accuracies for each classifier as opposed to a constant accuracy that would originate from the same topology of the data points.

In Fig. S1 is indicated that LDA is applied only for a number of components lower than the number of classes. Also, for a large variable-to-sample ratio, the shrinkage parameter, available in *scikit* LDA class, was set to "auto" to ensure a higher performance of LDA and thus a more fair comparison.

To express the statistical significance of the differences between accuracy distributions obtained with the algorithm outlined in Fig. S1, the Mann-Whitney *U* Test was applied. The null hypothesis (i.e. H0) states that the accuracy is not different from SuPP-X (X denotes SVM or NB, depending on the case) to the classifier Y (Y denotes any of the following: PCA-SVM, PCA-NB, PLS-DA and LDA) used for comparison. Alternative hypotheses are the following:

H1. The SuPP-X accuracy is higher than the accuracy of Y (right-tailed test)

H2. The SuPP-X accuracy is different than the accuracy of Y (two-tailed test)

H3. The SuPP-X accuracy is lower than the accuracy of Y (left-tailed test)

For the cases with the p-value bellow the significance level, the effect sizes (ES) are estimated following the work of Ruscio [45].

4. Results and discussion

4.1. Exploratory data analysis

For the assessment of the data visualization in lower dimensions using SuPP, a comparison with the separations obtained using PCA and PLS-DA was made. All three techniques were applied to the data set C described in the previous section. The results are illustrated in Fig. 4, where 2, 3 and 4 latent components of SuPP, PCA and PLS-DA were used to plot the data. From the 4D representation (Fig. 4 g, h and i) it is clear that, in case of the SuPP latent space, the groups are well separated in each of the 4 dimensions. This indicates a linear relationship between the latent components of SuPP and the classes in the data. The 2D representation in Fig. 4 a,b and c, indicates that there is a better separation of the classes in the case of SuPP (Fig. 4 a) while the distribution of the groups in the latent space is similar to those in the cases of PCA (Fig. 4 b) and PLS-DA (Fig. 4 c).

The pairwise distances and the statistical significance of the distances between pairs of classes represented in Fig. 4, are listed in Table 2. The

Table 2
Distances between pairs of groups of the projected "Wine" data set onto latent latent SuPP, PCA and PLS-DA spaces. The p-value is obtained from F-statistics and df1, df2° of freedom.

SuPP								
Group pair	MD	BD	HD	T2	F-statistics	p-val	df1	df2
1–2	1.8417	2.8525	0.9707	102.46	50.80	1.11E-16	2	119
2–3	1.8699	3.2160	0.9797	96.58	47.86	2.22E-16	2	111
1–2	1.8567	3.2876	0.9812	104.13	34.13	5.55E-16	3	118
2–3	1.8804	3.4684	0.9843	97.67	31.98	1.11E-15	3	110
1–2	1.8609	3.4056	0.9833	104.60	25.50	3.33E-15	4	117
2–3	1.8818	3.6409	0.9868	97.81	23.80	6.33E-15	4	109
PCA								
Group pair	MD	BD	HD	T2	F-statistics	p-val	df1	df2
1–2	1.7542	1.7641	0.9103	92.95	46.09	1.55E-15	2	119
2–3	1.8399	2.6621	0.9645	93.51	46.34	2.66E-15	2	111
1–2	1.7712	2.0102	0.9306	94.75	31.06	7.11E-15	3	118
2–3	1.8510	2.9012	0.9721	94.64	30.98	1.28E-14	3	110
1–2	1.7744	2.1503	0.9400	95.11	23.18	4.00E-14	4	117
	1.8510	3.1366	0.9780	94.64	23.03	7.02E-14	4	109
PLS								
Group pair	MD	BD	HD	T2	F-statistics	p-val	df1	df2
1–2	1.6935	1.3403	0.8592	86.63	42.95	9.10E-15	2	119
2–3	1.7734	1.9379	0.9252	86.87	43.05	1.53E-14	2	111
1–2	1.8249	2.5875	0.9617	100.59	32.97	1.44E-15	3	118
2–3	1.7919	2.1894	0.9423	88.69	29.04	2.78E-15	3	110
1–2	1.8386	2.8956	0.9720	102.10	24.89	6.33E-15	4	117
2–3	1.8540	2.9913	0.9746	94.95	23.10	1.19E-14	4	109

MD – Mahalanobis distance, BD - Bhattacharyya Distance, HD - Hellinger Distance, T2 – Hotelling's T^2, df1 – 1st parameter for degrees of freedom and latent components, df2 – 2nd parameter for degrees of freedom.

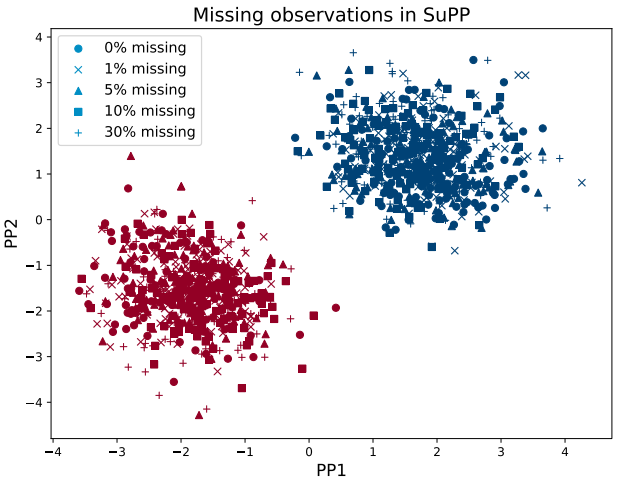

Fig. 5. The test set projected on the latent components obtained with different rates of missing observations.

dimension of the latent space is equivalent to the first parameter for the degrees of freedom, i.e. df1. The distances between the classes are significant for all three projection methods (SuPP, PCA and PLS-DA). However, the largest distances were obtained with SuPP (Table 2).

4.2. Optimization of the parameters

Detailed discussion on the optimization of the parameters of the SuPP and SVM is outlined in section 2.1 of the SI. From the results listed in Table S2.1 from SI, the penalty seems to play a small role in the accuracy of the prediction for both data sets when $D < 3$. Increasing D however, makes the effect of γ more evident. Thus, a recommendation is to explore the performance at γ values such as $\gamma = C^{-4}$ for a less orthogonal space and $\gamma = C^{6}$ for a more orthogonal space.

In Table S2.2 from SI, are listed the accuracy values of NB classifier and the average standard deviation of the latent components (i.e. "loadings"). The results are indicating a possible convergence of a solution for $\widehat{k}_1$ and existence of multiple solutions for $\widehat{k}_d$ with $d > 1$. In other words, the more iterations are used for higher dimensions, the more probable is to find different interesting projections. Additionally, Fig. S2.1 in SI, where the distributions of the correlation coefficient are plotted, supports this argument. The average accuracy of NB classifier however does not improve with the increased values of *niter*. Thus, a lower value, between 100 and 500, is recommended for a faster performance.

The optimization of the SVM classifier showed that for a low parsimony model and high accuracy, the combination of $C = 1$ and $\Gamma = 0.1$ is an optimal choice for both data sets A.II and A.III (Table S2.3).

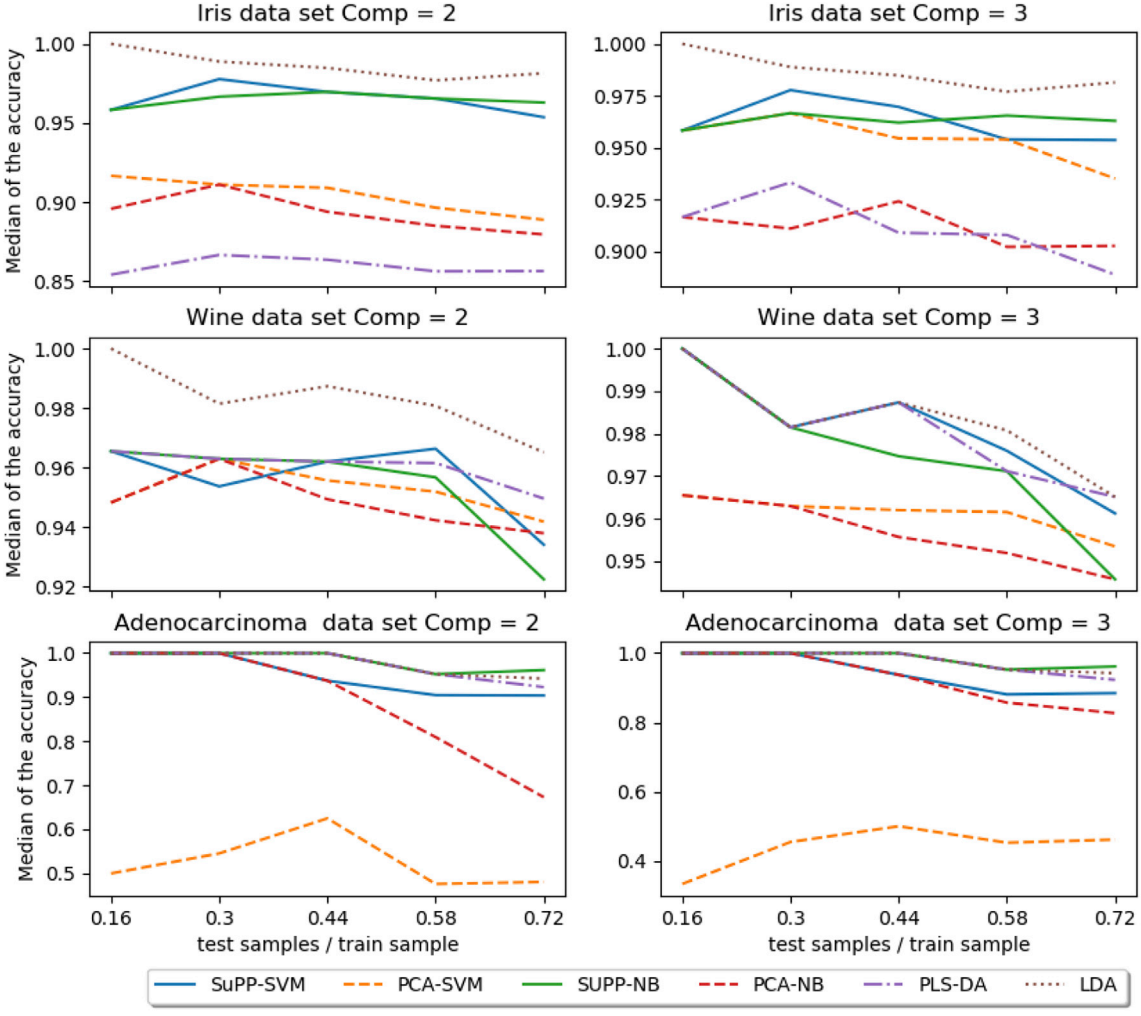

Fig. 6. The median of the accuracy distributions obtained using Cross-Validation strategy.

4.3. Handling missing data

As mentioned in the theory section using KDE of the projected data is an advantage when working with missing values (i.e. NA or NaN values). More specifically, the projection of the data is made without considering the missing data which will affect the histogram of the projection more or less significantly, depending on the amount and the location distribution of the missing data (Fig. 2). The estimation of the density in this case, plays a corrective (i.e. smoothing) role. The training data set included missing values in random location as described in section 3.3.3. The latent components were further used to project the test data set (Fig. 5). The effect of the missing data on the separation of the classes, in the latent space, was insignificant. This conclusion however must not be generalized to all data sets. The censoring of the data varies from case to case and as such, the effect of the missing values on the projection can sometimes be overwhelming (e.g. when the missing values are not uniformly distributed across the data set and moreover when the important features are missing in large proportions). These however are extreme cases and solutions for such cases are yet to be found.

4.4. SuPP as pre-processing for classification

The accuracy distributions obtained using CV algorithm outlined in Fig. S1, are represented in Figs. S4, S5 and S6 from the SI. The median of the accuracies, obtained with each of the methods used for comparison, are plotted in Fig. 6. The continuous lines correspond to SuPP.

For Iris data set the, SuPP is only superseded by LDA, for both 2 and 3 components. For Wine data set, the performance of SuP, in median accuracy, is at least as good as PLS-DA, for $D = 2$. While LDA is limited in the number of latent components, SuPP and PLS are not and this is reflected in the median accuracy trend for $D = 3$ in the case of Wine data set. In this case the median accuracy significantly increase for SuPP and PLS.

The p-values (of the alternative hypotheses) resulting from the hypotheses tests described in section 3.3.4 are presented in Table S4 to Table S7 from the SI. A simplified interpretation of Tables S4–S7 is the following: if there is a number on the H1 row, there is sufficient evidence that the classifier X, applied in the projected space obtained with SuPP, performs better than the corresponding classifier on column Y. If H2 row contains a number corresponding to a column Y, there is not sufficient evidence to reject the null hypothesis (i.e. the accuracy is about the same as for the classifier Y). Note that the p-values of this test are higher than the 5% significance level. And lastly, if there is a value on H3 row, corresponding to a column Y, the data provides sufficient evidence that the classifier Y performs better than SuPP-X. The assessment of the significance using p-values must be interpreted however with caution as this is not the most objective measure on its own. For p-values indicating significant improvement in accuracies from one method to another, an effect size (i.e. ES) value is attached (Table 3).

For the Wine data set (data set C), when the projection is made on two latent components (i.e. Table S4 in SI), SuPP-SVM does not perform better than any of the classifiers. In this case SuPP-SVM is less accurate than LDA. The same is observed in the case of SuPP-NB (Table S6 in SI). The performance changes when $D = 3$ for the same data set (Table S5 and Table S7) being at least as good or, as in the case of SuPP vs PCA, even better (p-value $\leq$ 2.95E-03 and ES $\geq$ 0.63). Small exceptions for $D = 3$ can be observed in Table S7 and Table S8 from the SI for test/training = 0.75 (Wine data set rows), where PLS-DA and LDA perform better than SuPP-NB. For the Wine data set the ES values are shown in Table S8.

For the Iris data set (data set B), with $D = 2$, SuPP performs better than PCA and PLS (p-value $\leq$ 4.02E-05 and p-value $\leq$ 1.17E-05 respectively) in all the increments of the test/training ratio. LDA however, in this case, performs better than SuPP. Table 3 indicates the p-values and

Table 3
The p-values and the effect sizes (ES) obtained by testing the hypothesis H1 for the Iris data set.

D = 2										
SuPP-SVM										
	test/training = 0.16		test/training = 0.3		test/training = 0.4		test/training = 0.58		test/training = 0.72	
	p-val	ES	p-val	ES	p-val	ES	p-val	ES	p-val	ES
PCA-SVM	7.84E-05	0.84	4.79E-07	0.95	2.27E-07	0.96	4.38E-08	0.99	3.68E-08	0.99
PLS-DA	4.28E-06	0.91	9.52E-08	0.98	8.61E-08	0.98	3.14E-08	0.998	3.67E-08	0.99
LDA[a]	0.92/8.7E-02	0.39	0.91/9.2E-02	0.38	0.99/5.8E-03	0.28	0.99/6.4E-04	0.27	0.99/7E-05	0.15
SuPP-NB										
	test/training = 0.16		test/training = 0.3		test/training = 0.4		test/training = 0.58		test/training = 0.72	
	p-val	ES	p-val	ES	p-val	ES	p-val	ES	p-val	ES
PCA-NB	4.02E-05	0.86	1.32E-07	0.97	5.06E-08	0.99	4.96E-08	0.99	4.63E-08	0.99
PLS-DA	1.17E-05	0.89	3.10E-07	0.96	6.84E-08	0.98	2.90E-08	1	3.68E-08	0.99
LDA[a]	0.98/2.1E-02	0.33	0.99/5.4E-03	0.27	0.99/4.4E-03	0.27	0.98/2E-02	0.31	0.99/2E-03	0.24
D = 3										
SuPP-SVM										
	test/training = 0.16		test/training = 0.3		test/training = 0.4		test/training = 0.58		test/training = 0.72	
	p-val	ES	p-val	ES	p-val	ES	p-val	ES	p-val	ES
PCA-SVM	5.92E-01	0.48	2.31E-01	0.56	4.87E-02	0.65	2.58E-01	0.56	4.09E-02	0.66
PLS-DA	2.92E-03	0.75	4.34E-05	0.85	4.28E-06	0.9	8.45E-06	0.89	1.60E-05	0.88
LDA[a]	0.92/8E-02	0.38	0.99/5.1E-03	0.27	0.99/2.7E-03	0.25	0.99/9.9E-05	1.60E-01	0.99/1.1E-04	0.16
SuPP-NB										
	test/training = 0.16		test/training = 0.3		test/training = 0.4		test/training = 0.58		test/training = 0.72	
	p-val	ES	p-val	ES	p-val	ES	p-val	ES	p-val	ES
PCA-NB	3.00E-03	0.74	1.66E-04	0.83	6.56E-05	0.85	5.92E-06	0.902	2.39E-06	0.92
PLS-DA	3.08E-03	0.74	1.77E-04	0.82	1.23E-05	0.89	2.98E-07	0.96	1.17E-06	0.93
LDA[a]	0.95/5.5E-02	0.36	0.99/2.4E-03	0.25	0.99/1.6E-03	0.23	0.99/1.3E-03	0.23	0.99/4.3E-03	0.26

[a] For LDA, the values after the slash represent the p-values of the alternative hypothesis "The accuracy of SuPP-X is below LDA".

the ES values for the Iris data set when testing the hypothesis H1 defined previously. A significantly large effect size in this cases (ES $\geq$ 0.84 and ES $\geq$ 0.89 for PCA and PLS respectively) indicates a better separation of the classes in latent spaces. For D = 3 the application on Iris data set shows better performance than PLS-DA (p-value $\leq$ 3.08E-03 and ES $\geq$ 0.74) and, with small exceptions (Table S5 test/training = 0.16, 0.3 and 0.58), better than PCA (p-value $\leq$ 3.00E-03). The ES values however for this data set, for D = 3, are lower than the correspondent ones for a 2D latent space.

For data set D (adenocarcinoma), SuPP-SVM performed better with respect to PCA-SVM (Table S4). Compared to PLS-DA and LDA, for D = 2 the accuracy of SuPP-SVM is at least as good or lower. Increasing the dimensions of the latent space did not improve the performance of SuPP-SVM on the adenocarcinoma data set. For SuPP-NB, the accuracy is at least as good as for the other classifiers and in some cases (test/training $\geq$ 0.44) SuPP-NB performs better than PCA-NB (Table S2.6 and Table S7). The results from Tables S6 and S7 point to the optimality of SuPP as a preprocessing or dimensionality reduction step applied prior to a probabilistic classifier (in this case NB) as indicated in eq. (5). The ES values for the data set D are included in Table S10.

5. Conclusions

The SuPP strategy described here is a versatile dimensionality reduction technique that offers a new perspective on supervised exploratory data analysis. SuPP proved to be able to work for low samples-to-variables ratio as well as with missing values from the data set. The main theoretical aspect of this method indicates that JSD is an important metric for a comprehensive representation in lower dimensions. The method, although slower (i.e. 13 min for A.II, D = 2, niter = 100 and 23 min for A.II, D = 3, niter = 100) than the classical dimensionality reduction techniques like PCA and PLS, is more objective (i.e. less prone to overfitting) and in some cases a better choice for preliminary step in classification process. The superiority of SuPP was observed especially on the Iris data set, where the accuracy of SuPP-NB and SuPP-SVM were above PCA-NB, PCA-SVM and PLS-DA. In the case of the Wine data set, SuPP performs at least as good as LDA for 50% training-to-test sample ratio. The capacity of SuPP to reduce the dimension independently of the parameters of the projected distribution, could, potentially show better result than LDA on a non-normally distributed projected distributions.

SuPP is an alternative to the classical DR methods and the choice to use SuPP must be determined by the data itself and the desire of the user to discover new perspectives on the projected data. Future work will include a feature selection approach based on SuPP, applications on categorical data and a more extensive application on the biological data sets.

Fundings

This study was supported by the Human Nutrition & Health initiative of the University of Groningen.

Declaration of competing interest

None declared.

Acknowledgements

The author would like to acknowledge Prof. Dr. Folkert Kuipers and Prof. Dr. Ido Kema for valuable contribution to the initiative of the Nutrition & Health of the RUG.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.chemolab.2019.103867.

References

[1] K. Pearson, On lines and planes of closest fit to systems of points in space, Philos. Mag. 2 (11) (1901) 559–572.
[2] H. Hotelling, Analysis of a complex of statistical variables into principal components, J. Educ. Psychol. 24 (6) (1933) 417–441.
[3] J.H. Friedman, J. Tukey, A projection pursuit algorithm for exploratory data analysis, IEEE Trans. Comput. C-23 (9) (September 1974) 881–890.
[4] J.H. Friedman, W. Stuetzle, Projection pursuit regression, J. Am. Stat. Assoc. 76 (376) (1981) 817–823.
[5] Q. Guo, W. Wu, F. Questier, D.L. Massart, Sequential projection pursuit using genetic algorithms for data mining of analytical data, Anal. Chem. 72 (13) (2000) 2846–2855.
[6] A. Hyvärinen, J. Karhunen, E. Oja, Independent Component Analysis, Wiley, New York: NY, 2001.
[7] J.B. Tenenbaum, V. de Silva, J.C. Langford, A global geometric framework for nonlinear dimensionality reduction, Science 290 (5500) (2000) 2319–2323.
[8] C. Hou, C. Zhang, Y. Wu, F. Nie, Multiple view semi-supervised dimensionality reduction, Pattern Recognit. 43 (3) (2010) 720–730.
[9] K.Ø. Mikalsen, C. Soguero-Ruiz, F.M. Bianchi, Noisy multi-label semi-supervised dimensionality reduction, Pattern Recognit. 90 (2019) 257–270.
[10] H. Wold, Estimation of principal components and related models by iterative least squares, Multivar. Anal. (1966) 391–420.
[11] H. Wold, Partial least squares, Encycl. Stat. Sci. 6 (1985) 581–591.
[12] R.A. Fisher, The use of multiple measurements in taxonomic problems, Ann. Eugen. 7 (2) (1936) 179–188.
[13] T. Hastie, R. Tibshirani, Discriminant analysis by Gaussian mixtures, J. R. Stat. Soc. Ser. B 58 (1) (1996) 155–176.
[14] D.-Q. Dai, P. Yuen, Face recognition by regularized discriminant analysis, IEEE Trans. Syst. Man Cybern. B Cybern. 37 (7) (2007) 1080–1085.
[15] K.-K. Huang, D.-Q. Dai, C.-X. Ren, Regularized coplanar discriminant analysis for dimensionality reduction, Pattern Recognit. 62 (2017) 87–98.
[16] C. Örnek, E. Vural, Nonlinear supervised dimensionality reduction via smooth regular embeddings, Pattern Recognit. 87 (2019) 55–66.
[17] B. Raducanu, F. Dornika, A supervised non-linear dimensionality reduction approach for manifold learning, Pattern Recognit. 45 (2012) 2432–2444.
[18] A.M. Pires, Robust linear discriminant analysis and the projection pursuit approach, Dev. Robust Stat. (2003) 317–329.
[19] A.M. Pires, J.A. Branco, Projection-pursuit approach to robust linear discriminant analysis, J. Multivar. Anal. 101 (10) (2010) 2464–2485.
[20] E.-K. Lee, D. Cook, S. Klinke, T. Lumley, Projection pursuit methods for exploratory supervised classification, J. Comput. Graph. Stat. 14 (4) (2005) 831–846.
[21] S. Buchala, N. Davey, T.M. Gale, R.J. Frank, Analysis of linear and nonlinear dimensionality reduction methods for gender classification of face images, Int. J. Syst. Sci. 36 (14) (2005) 931–942.
[22] A. Konstorum, N. Jekel, E. Vidal, R. Laubenbacher, Comparative Analysis of Linear and Nonlinear Dimension Reduction Techniques on Mass Cytometry Data, 2018, pp. 1–16, bioRxiv.
[23] A. Akhbardeh, M.A. Jacobs, Comparative analysis of nonlinear dimensionality reduction techniques for breast MRI segmentation, Med. Phys. 39 (4) (2012) 2275–2289.
[24] A. Errity, J. McKenna, A comparative study of linear and nonlinear dimensionality reduction for speaker identification, in: 15th International Conference on Digital Signal Processing, Cardiff, 2007, 2007.
[25] K.T. Abou-Moustafa, F. De La Torre, F.P. Ferrie, Pareto models for discriminative multiclass linear dimensionality reduction, Pattern Recognit. 48 (2015) 1863–1877.
[26] P.S. Gromski, H. Muhamadali, D.I. Ellis, Y. Xu, E. Correa, M.L. Turner, R. Goodacre, A tutorial review: metabolomics and partial least squares-discriminative analysis - a marriage of convenience or a shotgun wedding, Analytica Chemica Acta 879 (2015) 10–23.
[27] J.A. Westerhuis, H.C.J. Hoefsloot, S. Smit, D.J. Vis, A.K. Smilde, E.J. van Velzen, J.P.M. van Duijnhoven, F.A. van Dorsten, Assessment of PLSDA cross validation, Metabolomics 4 (1) (2008) 81–88.
[28] J. Kruskal, Toward a practical method which helps uncover the structure of a set of observations by finding the line transformation which optimizes a new "index of condensation, Stat. Comput. (1969) 427–440.
[29] P.J. Huber, Projection pursuit, Ann. Stat. 13 (2) (1985) 435–475.
[30] M. Jones, R. Sibson, What is projection pursuit? J. R. Stat. Soc. Ser. A 150 (1) (1987) 1–37.
[31] A. Rényi, On measures of information and entropy, in: Proceedings of the Fourth Berkeley Symposium on Mathematics, Statistics and Probability, 1960, pp. 547–561, 1961.
[32] B. Fuglede, F. Topsoe, Jensen-Shannon divergence and Hilbert space embedding, in: International Symposium onInformation Theory, 2004. ISIT 2004. Proceedings. vol. 31, 2004.
[33] J. Briët, P. Harremoës, Properties of classical and quantum Jensen-Shannon divergence, Phys. Rev. A 79 (5) (2009).
[34] J. Lin, Divergence measures based on the Shannon entropy, IEEE Trans. Inf. Theory 37 (1) (1991) 145–151.
[35] S. Hou, P.D. Wentzell, Regularized projection pursuit for data with a small sample-to-variable ratio, Metabolomics 10 (2014) 589–606.
[36] L. Zhenqin, H. Sharega, Monte Carlo-minimization approach to the multiple-minima problem in protein folding, Proc. Natl. Acad. Sci. 84 (1987) 6611–6615.
[37] M. Powel, in: S. Gomez, J. Hennart (Eds.), A Dierct Search Optimization Method that Models the Objective and Constraint Functions by Linear Interpolation, vol. 275, Springer, Dordrecht, 1994, pp. 51–67.

[38] T.E. Oliphant, Python for scientific computing, Comput. Sci. Eng. 9 (2007) 10–20.
[39] E. Anderson, The species problem in Iris, Ann. Mo. Bot. Gard. 23 (3) (1936) 457–509.
[40] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.
[41] M. Forina, R. Leardi, C. Armanino, S. Lanteri, PARVUS: an Extendable Package of Programs for Data Exploration, Classification and Correlation, 1988.
[42] D. Notterman, U. Alon, A. Sierk, A. Levine, Transcriptional gene expression profiles of colorectal adenoma, adenocarcinoma, and normal tissue examined by oligonucleotide arrays, Cancer Res. 61 (7) (2001) 3124–3130.
[43] X. Shao, H. Li, N. Wang, Q. Zhang, Comparison of different classification methods for analyzing electronic nose data to characterize sesame oils and blends, Sensors 15 (10) (2015) 26726–26742.
[44] K. Roy, P. Ambure, The "double cross-validation" software tool for MLR QSAR model development, Chemometr. Intell. Lab. Syst. 159 (2016) 108–126.
[45] J. Ruscio, A probability-based measure of effect size: robustness to base rates and other factors, Psychol. Methods 13 (1) (2008) 19–30.